① LINEAR REGRESSION - EdX AAnalytics

One variable regression model

- $y_i = \beta_0 + \beta_1 x^i + \epsilon^i$

y_i = dependant variable for the i^{th} observation
x_i = independant variable for the i^{th} observation
ϵ^i = error term for the i^{th} observation (epsilon)
β_0 = intercept coefficient
β_1 = regression coefficient for independant variable
- The best model (choice of coefficients) has the smallest error terms/residuals.

Goal of linear regression = create a predictive line.

- error = residual = actual value of a data point - prediction
- SSE (Sum of squared errors) = measure the quality of a regression line

$SSE = (\epsilon^1)^2 + (\epsilon^2)^2 + \dots + (\epsilon^n)^2$ n = nb of data points

. can be hard to interpret = depends on n & units are hard to understand
- RMSE (Root-Mean-Squared Error) $= \sqrt{\frac{SSE}{n}}$

. normalized by n, units of dependant variable
- R^2 = Compares the best model to a "baseline" model (baseline model: predicts the average value of the dependant variable regardless of the value of the independent variable).

② LINEAR REGRESSION - Edx Analytics

SST = sum of squared errors for the baseline model

- $R^2 = 1 - \frac{SSE}{SST}$

$0 \leq SSE \leq SST$

- R^2 captures value added from using a model
  - $R^2 = 0$ means no improvement over baseline
  - $R^2 = 1$ means a perfect predictive model
- Unitless & universally interpretable
  - Can still be hard to compare between problems
  - Good models for easy problems will have $R^2 \approx 1$
  - Good models for hard problems can still have $R^2 \approx 0$

Multiple linear regression model with k variable

$$y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots + \beta_k x_k^i + \epsilon^i$$

y^i = dependent variable for the i^{th} observation
x_j^i = j^{th} independent variable for the i^{th} observation
ϵ^i = error term for the i^{th} observation
β_0 = intercept coefficient
β_j = regression coefficient for the j^{th} independent variable

- Best model coefficients selected to minimize SSE.

⇒ Adding variable

- Adding more variables can improve the model
- Diminishing returns as more variables are added

(3) LINEAR REGRESSION - EdX AAnalytics

- Not all available variables should be used
  - each new variable requires more data
  - Causes overfitting = high R^2 on data used to create model but bad performance on unseen data
- The model's R^2 value can never decrease from adding new variables to the model.

Correlation: a measure of the linear relationship between variables.

+1 = perfect positive linear relationship
0 = no linear relationship
-1 = perfect negative linear relationship

Two variable highly correlated: absolute value of correlation close to 1.
Typically, a correlation greater than 0.7 or -0.7 is cause for concern.

Out of sample accuracy: accuracy of the model on the test data.

(4) LOGISTIC REGRESSION - Edx AA

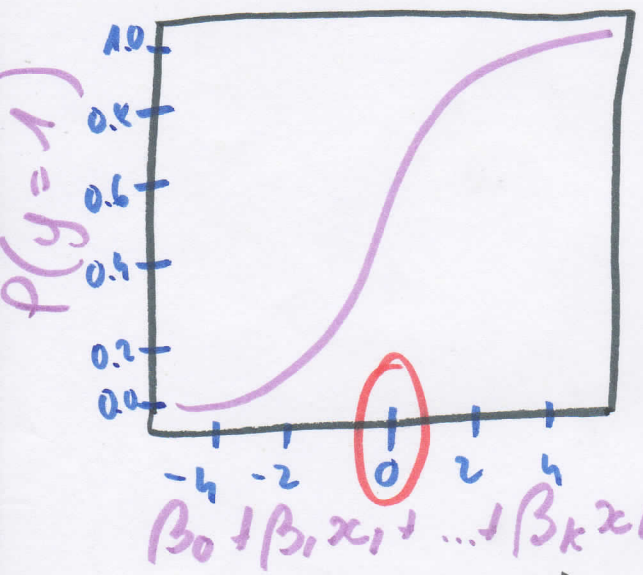
Logistic regression: predict the probability of the outcome variable being true $P(y=1) = 1 - P(y=0)$

The outcome is either 1 or 0. (binary categorical)

Independent variables: $x_1, x_2, x_3 \dots, x_k$

Logistic Response function: used to produce a number between 0 and 1

$$P(y=1): \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_k x_k)}}$$



- Positive values are predictive of class 1.
- Negative values are predictive of class 0.

$$\text{Odds} = \frac{P(y=1)}{P(y=0)}$$

Odds > 1, if $y=1$ is more likely

Odds < 1, if $y=0$ is more likely

$$\text{Odds} = e^{\beta_0+\beta_1 x_1+\dots+\beta_k x_k}$$

$$\log(\text{Odds}) = \beta_0+\beta_1 x_1+\dots+\beta_k x_k = \text{"logit"}$$

The bigger the logit is, the bigger $P(y=1)$

The outcome of a logistic regression model is a probability of being true

(4) LOGISTIC REGRESSION - Edx AA

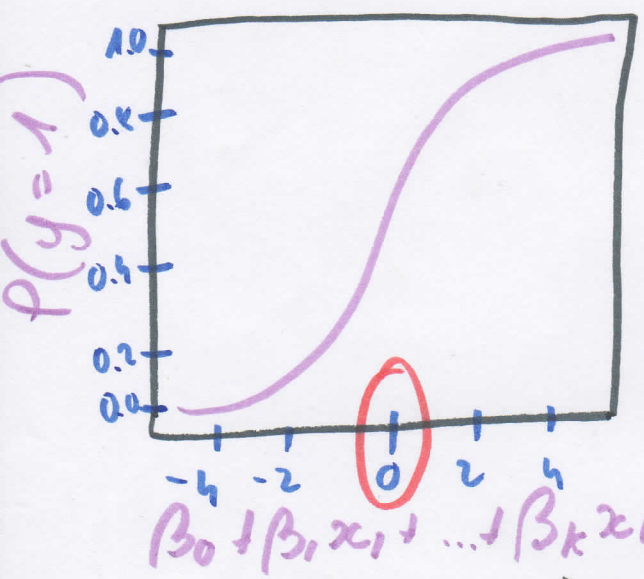
Logistic regression: predicts the probability of the outcome variable being true $P(y=1) = 1 - P(y=0)$

The outcome is either 1 or 0. (binary categorical)

Independent variables: $x_1, x_2, x_3 \dots, x_k$

Logistic Response function: used to produce a number between 0 and 1

$$P(y=1): \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_k x_k)}}$$



- Positive values are predictive of class 1.
- Negative values are predictive of class 0.

$$\text{Odds} = \frac{P(y=1)}{P(y=0)}$$

Odds > 1, if $y=1$ is more likely

Odds < 1, if $y=0$ is more likely

$$\text{Odds} = e^{\beta_0+\beta_1 x_1+\dots+\beta_k x_k}$$

$$\log(\text{Odds}) = \beta_0+\beta_1 x_1+\dots+\beta_k x_k = \text{"logit"}$$

The bigger the logit is, the bigger $P(y=1)$

The outcome of a logistic regression model is a probability of being true

⑤ LOGISTIC REGRESSION Edx AA

Confusion matrix (classification matrix)

Compare actual outcome to predicted outcome

	Predicted = 0	Predicted = 1
Actual = 0	True negative: TN	False Positive: FP
Actual = 1	False negative: FN	True positive: TP

Sensitivity = $\frac{TP}{TP+FN}$ (True positive rate)

% of actual = 1 classified correctly

Measure the

Specificity = $\frac{TN}{TN+FP}$ (True negative rate)

Measure the % of actual = 0 classified correctly

Model with higher threshold = lower sensitivity and higher specificity

With lower threshold: higher sensitivity and lower specificity

The threshold transform the probability outcome into a binary outcome

Choose best threshold for best trade off:

- cost of failing to detect positives
- cost of raising false alarms

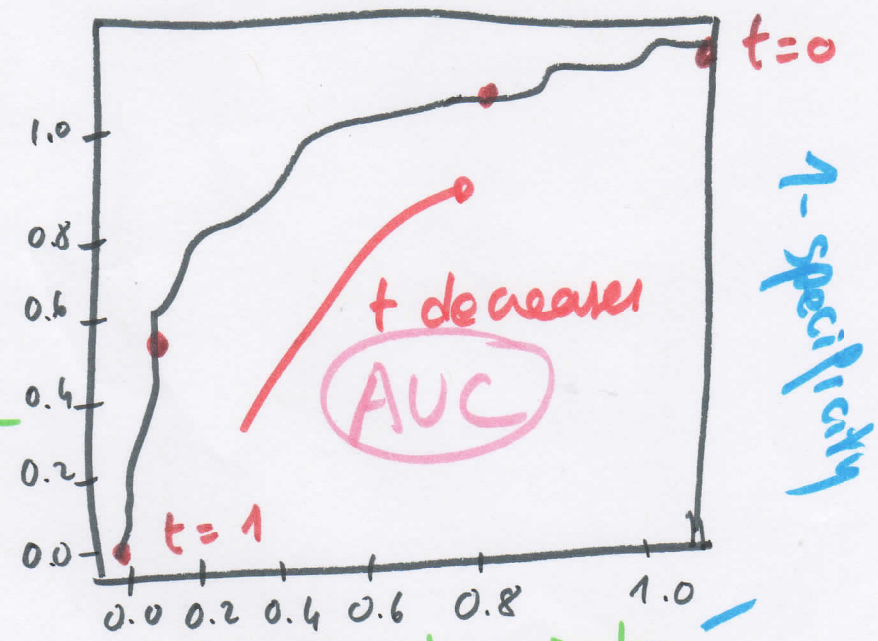
Use of threshold value t to make a binary prediction

if $P(HIV=1) \geq t$, predict HIV | if $P(HIV=1) < t$ predict no HIV

6) LOGISTIC REGRESSION EdX AA

Receiver Operator Characteristic (ROC) Curve

- True Positive rate (sensitivity) on y-axis
- False Positive rate (1 - specificity) on x-axis
- t = threshold



t = 1 (threshold = 1) sensitivity of 0 & false positive rate of 0. Not catch the actual = 1, but catch all actual = 0

t = 0 sensitivity of 1 & false positive rate = 1. Catch all actual = 1 but label all actual = 0 as 1

Point (0, 0.4): label 40% of actual = 1 with small false positive rate

Point (0.6, 0.9): correctly labelling 90% of actual = 1 but false positive rate of 60%.

- The Roc curve captures all thresholds simultaneously

- High threshold
  - High specificity
  - Low sensitivity

- Low threshold
  - low specificity
  - high sensitivity

(Area Under the Curve)

AUC = gives an absolute measure on how good is the model

⑦ LOGISTIC REGRESSION EdX AA

n = number of observations

Overall accuracy = $\frac{(TN+TP)}{n}$ (suite page 13)

Overall error rate : $(FP+FN)/n$

False Negative error rate : $FN/(TP+FN)$

False Positive Error rate : $FP/(TN+FP)$

TREES

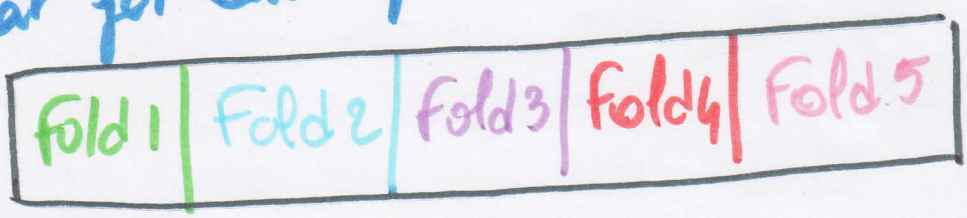
CART : Classification and Regression Trees

We also use ROC curve

The value of minbucket can affect the model's out of sample accuracy. If it is too small, overfitting might occur. If it is too large, the model might be too simple (rpart(..., minbucket = ..))

k-fold cross validation

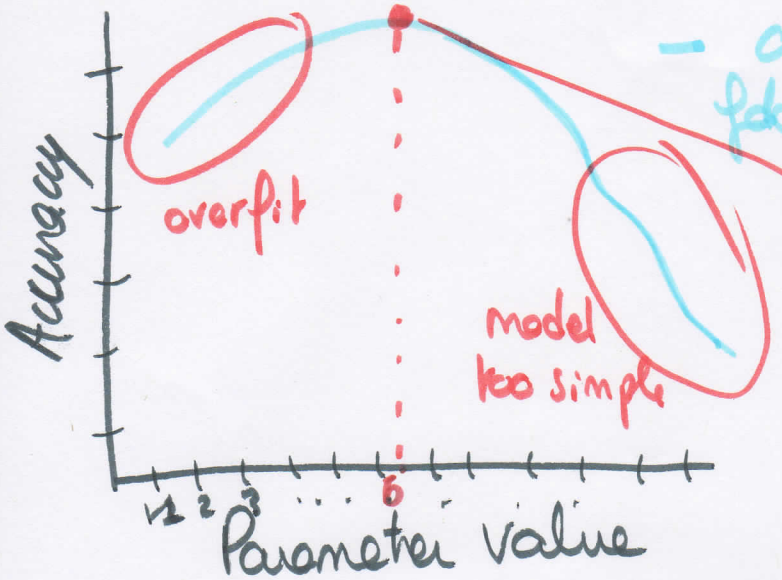
- Given training set, split into k pieces (here k = 5)
- Use k-1 folds to estimate a model, and test model on remaining one fold ("validation set") for each candidate parameter value
- Repeat for each of the k-folds



) training set

(8) TREES EdX AA

Output of k-fold cross-validation



- average of output from fold 1, 2, 3, 4 & 5 (all)

maximum average accuracy over all parameter value in R, parameter is called "cp" (complexity parameter)

Penalty error / matrix

- Use asymmetric penalties
- Define a "penalty matrix" as the cost of being wrong

Forecast \ Outcome	1	2	3	4	5	6
1	0	2	3	4	5	6
2	1	0	2	3	4	5
3	2	1	0	2	3	4
4	3	2	1	0	2	3
5	4	3	2	1	0	2
6	5	4	3	2	1	0

- To judge the quality of the analytics models, we compare it with a **baseline** (predict that the cost in the next period will be the cost in the current period.

TREES

"Cp" parameter = complexity parameter

- Having too many splits is bad for generalization, so we should penalize the complexity
- RSS: Residual sum of squares (sum of the square $\neq$s)

$$\sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Goal: minimize the RSS by making splits but penalize too many splits. S: number of splits

$$\sum_{\text{leaves}} (\text{RSS at each leaf}) + \lambda S \qquad \lambda = \text{penalty}$$

Tree that minimizes

- λ large: not many splits because we pay a big price for every additional split outweighting the decrease in "error"
- λ small or zero: make splits until it no longer decreases error
- $cp = \dfrac{\lambda}{RSS(\text{no splits})}$
  - cp small: encourage large trees
  - cp large: encourage small trees

TEXT ANALYTICS

Bag of words

- Count the number of times each word appears
- Simple approach but effective. Used as a baseline in text analytics projects & natural language processing.

(10) TEXT ANALYTICS Edx AA

- Preprocessing the text can dramatically improve the performance of bag of words method.
  - → change all words to either lower-case or upper-case
  - → Remove everything that isn't a standard number or letter (⚠ sometimes punctuation is meaningful)
  - → Remove unhelpful terms (stop words = the, is, at, which..)
  - ⚠ two stop words taken together might have important meaning. Ex: The Who (music band)
  - → Stemming (racine) ① build a database of words and their stems Handle exceptions. Con: won't handle new words ② "Porter Stemmer" (rule-based algorithm) ex: if word ends in "ed", "ing" remove it. Handles new/unknown words. Con: many exceptions

Text data preprocessing steps example

packages "tm" and "SnowballC"

1. Create corpus
2. Convert to lower case
3. Convert to plain text document
4. Remove punctuation
5. Remove stopwords
6. Stem document
7. Remove sparse terms (create matrix with frequencies of words in doc)
8. Convert to data frame
9. Add dependent variable (outcome)
10. Split data (train & test), build models, test accuracy, compare baseline

(11) CLUSTERING EdX AA

- Unsupervised learning
- Goal: segment the data into similar groups instead of prediction
- Cluster data into 'similar' groups and then build a predictive model for each group
- ⚠ Overfitting. Works best with large datasets
- Many different algorithms for clustering
  => Most popular: ① Hierarchical ② k-means

Distance between points - Compute distance

- Most popular: "Euclidean distance"

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2}$$

where k is the number of independent variables

- Minimum distance: distance between points of clusters that are the closest
- Maximum distance: distance between points of clusters that are the farthest
- Centroid distance: distance between centroids of clusters. Centroid: point that has the average of all data points in each component. Most used

Meaningful clusters? look at statistics (mean, min, max..) for each cluster and each variable

(12)

CLUSTERING

Hierarchical clustering

① Start with each data point in its own cluster
② Combine two nearest clusters (Euclidean, Centroid) until one single cluster
③ Choose how many clusters (look dendrogram. Height of vertical line represents distance between points or clusters)
④ Meaningful cluster? Look at statistics (mean, min, max...) for each cluster and each variable + see if the clusters have a feature in common that was not used in the clustering (like an outcome)

K-means clustering

- Aims at partitioning the data into k clusters in which each data point belongs to the cluster whose mean is the nearest
- **Algorithm:**

① Specify desired number of clusters k
② Randomly assign each data point to a cluster
③ Compute cluster centroids
④ Re-assign each point to the closest cluster centroid
⑤ Re-compute cluster centroids
⑥ Repeat 4 and 5 until no improvement is made.

(13)

CLUSTERING

- Cluster are interpretable and reveal unique patterns of diagnostic history among the population.

VISUALIZATION

- One picture is worth a thousand words

Useful for:

① Data exploration (better understand the data)
② Interpreting a model
③ Communicating the results

- **Data visualization**: a mapping of data properties to visual properties.

Data properties: usually numerical or categorical
Visual properties: can be (x, y) coordinates, colors, sizes, shapes, heights...

- Visualization can bridge the gap between the analytics and the end user.

AUC (Area Under the Curve): Gives an absolute measure of quality of the model. Shows how accurate the model is in an absolute sense

AUC = 0.5 random guessing AUC above 0.8 "good"
AUC = 1 perfect classifier

The threshold transforms the probability outcome of logistic regression into a binary outcome ⚠